A MATHEMATICAL TREATMENT OF SOME BIOLOGICAL PROBLEMS.

SHINKISHI HATAI,

ASSOCIATE IN NEUROLOGY, THE WISTAR INSTITUTE OF ANATOMY.

Dr. King's work ('07, '09), "On Studies on Sex-determination in Amphibians," suggests an interesting problem which may be put in the following form:

A jar contains a large number of male and female tadpoles, the proportion of each being unknown: if on picking out $m+n$ tadpoles, m are found to be males and n females, to find the probability that the ratio of the number of either sex to the entire lot lies between given limits.

It will be seen that problems of this nature occur frequently in biological investigations and that it is of importance to have some method for determining the accuracy of the observed proportions. This can be done by means of the formula given below. As the development of the formula is somewhat complicated I shall present the entire process of the mathematical treatment of the solutions based on the theorem of Bayes, together with one application. Although such an elementary exposition of the subject will be superfluous for one who is familiar with the theory of probabilities nevertheless for others it may be helpful.

If p denote the probability that an event will happen, then $(1-p)$ is the probability that the event will fail. If the probability that the event will fail on any single trial is $(1-p)$, the probability that it will fail every time is $(1-p)^n$. The probability that it will happen on the first trial and fail on the succeeding $n-1$ trial is $p(1-p)^{n-1}$. But the event is just as likely to happen on the second, third, etc., trials as on the first. Hence the probability that the event will happen just once in the n trials is

$$np(1-p)^{n-1}.$$

Continuing this process, we can easily see that the probability that it will happen m times in $m+n$ trials is

$$\frac{|\underline{m+n}}{|\underline{m}\,|\underline{n}}\,p^m(1-p)^n. \tag{1}$$

From (1) we can obtain directly the probability that exactly m males and n females occur in $m+n$ drawings and the expression will be

$$p = \frac{|\underline{m+n}}{|\underline{m}\,|\underline{n}}\,x(1-x)^n \tag{2}$$

where x denotes males and $(1-x)$ females.

If we call the right hand member of (2) y, then

$$y = cx^m(1-x)^n$$

and this equation represents a curve with a zero ordinate when $x = 0$ and $x = 1$. Since the elementary area under the curve is

$$\frac{|\underline{m+n}}{|\underline{m}\,|\underline{n}}\,x^m(1-x)^n dx,$$

the number of cases which occur between the two limits (a, b) is proportional to the sum of the differentials taken from a to b, or

$$\frac{|\underline{m+n}}{|\underline{m}\,|\underline{n}}\int_a^b x^m(1-x)^n dx.$$

Also the totality of cases will be proportional to the sum of the differentials taken from 0 to 1, or

$$\frac{|\underline{m+n}}{|\underline{m}\,|\underline{n}}\int_0^1 x^m(1-x)^n dx.$$

Hence the probability that the ratio of the males in the jar to the entire lot lies between the two limits is

$$p = \frac{\int_a^b x^m(1-x)^n dx}{\int_0^1 x^m(1-x)^n dx}. \tag{3}$$

This is the well-known theorem of Bayes.[1] This equation as it

[1] See Todhunter's "History of the Theory of Probability from the Time of Pascal to that of Laplace," 1865.

stands is applicable to cases where the number of observations is small. For cases where the number of observations is large we must modify this still further.

If we suppose the two limits to be $m/s \pm \theta$ where $s = m + n$, then equation (3) may be written in the following form :

$$p = \frac{\int_{m/s-\theta}^{m/s+\theta} x^m(1-x)^n dx}{\int_0^1 x^m(1-x)^n dx}.$$

By successive integration by parts the denominator is evaluated, giving

$$\int_0^1 x^m(1-x)^n dx = \frac{m\,!\,n\,!}{(m+n+1)\,!}. \tag{4}$$

If we let $x = m/s + z$ the numerator becomes

$$\int_{-\theta} \left(\frac{m}{s} + z\right)^m \left(\frac{n}{s} - z\right)^n dz,$$

which is approximately

$$\frac{m^m n^n}{s^s} \int_{-\theta}^{+\theta} \varepsilon^{-\frac{s^3 z^2}{2mn}} dz.$$

With the above transformation the formula becomes

$$p = \frac{(s+1)\,!}{m\,!\,n\,!} \frac{m^m n^n}{s^s} \int_{-\theta}^{+\theta} \varepsilon^{-\frac{s^3 z^2}{2mn}} dz. \tag{5}$$

Now if we apply Stirling's formula for large numbers (4) becomes

$$\frac{(s+1)\,!}{m\,!\,n\,!} = \frac{s^s}{m^m n^n} \sqrt{\frac{s^3}{2\pi mn}}.$$

Therefore (5) may be written in the following form :

$$p = \sqrt{\frac{s^3}{2\pi mn}} \int_{-\theta}^{+\theta} \varepsilon^{-\frac{s^3 z^2}{2mn}} dz.$$

If we assume

$$\sqrt{\frac{s^3}{2mn}}\, z = t$$

and substitute ω for

$$\theta\sqrt{\frac{s^3}{2mn}}$$

then as a final form we have

$$p = \frac{2}{\sqrt{\pi}}\int^{\omega} \varepsilon^{-t^2}dt.$$

Since the above expression is a well-known equation for the probability integral, the degree of probability for any given limits may be determined from the table. For computation we have the following relation :

$$\omega = \theta\sqrt{\frac{s^3}{2mn}} \quad \text{or} \quad \theta = \omega\sqrt{\frac{2mn}{s^3}},$$

and $$p = \varphi(\omega).$$

As will be seen from the above relations, the degree of probability is proportional to the amount of deviation. Therefore in any given case we can either increase or decrease the value of probability by changing the value of the deviation. Thus in order to facilitate a comparison of several sets of data, it would be advantageous to fix the value of the probability and then determine the corresponding amount of the deviation. If we take 0.75 for the value of the probability, it will be certainly high enough for practical purposes. If we adopt this system, then the corresponding value of ω is fixed and is equal to 0.814. Thus we do not even need to use the table to determine the amount of deviation. If however one wishes to find any other value of the probability than I have proposed (*i. e.*, 0.75), such can be readily obtained from the table. Thus the determination of the deviation can be made by a simple arithmetical process.

The following will illustrate a method of determination. Dr. King has kindly supplied the data for this purpose and I wish to thank her for that material.

Example : Out of 16,100 tadpoles, 9,949 are examined. 5,136 are found to be females and 4,813 males. Find the probability that the ratio of females to the entire lot lies between given limits.

We have

$$\theta = \omega \sqrt{\frac{2mn}{s^3}}. \qquad p = \varphi(\omega) = 0.75.$$

Since we adopted the fixed value for probability (0.75) the value of ω is also fixed and equal to 0.814. Thus our problem is therefore to find a value of θ when the value of p is known.

$$\theta = 0.814 \sqrt{\frac{2 \times 4{,}813 \times 5{,}136}{99{,}493}} = 0.814 \times 0.00709 = 0.0058.$$

Deviation = $16{,}100 \times 0.0058 = \pm 93$.

Total number of females = $16{,}100 \times 5{,}136/9{,}949 = 8{,}311 \pm 93$.

With this value of probability we conclude that the total number of females in the entire lot is neither greater than 8,404 nor less than 8,218.